THE PSYCHOLOGICAL CORPORATION

and the

U. S. NAVAL SCHOOL OF AVIATION MEDICINE

Joint Research Report

An evaluation of an experimental flight grading method for use in the U. S. Naval Air Basic Training Command.

REX D. DANNESKIOLD

and

LCDR WOODBURY JOHNSON, MSC, USN

THE PSYCHOLOGICAL CORPORATION

JOINT PROJECT REPORT

The Psychological Corporation, New York, New York
under Contract Nonr-1162(00)
and
U. S. Naval School of Aviation Medicine
Project No. NM 001 058.24.03

AN EVALUATION OF AN EXPERIMENTAL FLIGHT GRADING METHOD FOR USE IN THE NAVAL AIR BASIC TRAINING COMMAND

Report by

Rex D. Danneskiold

and

Lieutenant Commander Woodbury Johnson, MSC, USN

Approved by

George K. Bennett, President
The Psychological Corporation
and
Captain Ashton Graybiel, MC, USN
Director of Research
U. S. Naval School of Aviation Medicine

23 August 1954

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The authors wish to express their indebtedness to the Staff Instructors of the Instructor Basic Training Unit (Flight), U. S. Naval Air Station, Pensacola, Florida, who participated in the development and tryouts of these grading forms. Individual acknowledgment is made to Commander Philip C. Morris, USN and Lieutenant Commander Glen I. Palmer, USNR, whose interest and cooperation made it possible for this study to be carried on. Appreciation is also expressed to the members of the Pilot Advisory Board whose contributions to the development of these forms was invaluable. They are Major Lloyd J. Parsons, USMC, Lieutenant (jg) Clyde W. Summit, USN, and Lieutenant (jg) Edward P. Kellogg, USNR.

Acknowledgment is made for the counsel of Dr. John T. Bair, and the statistical advise of Dr. Marshall B. Jones, both of the U. S. Naval School of Aviation Medicine.

Appreciation is also expressed to Mrs. Margaret Crawley, Mrs. Claire Danneskiold and Mrs. Betty Wilson for their efforts in the statistical, clerical and secretarial aspects of the project.

SUMMARY

As an outgrowth of earlier research in Navy flight grading (8, 12), this study evaluated a basically different type of daily flight grading form from that currently in use in the Naval Air Basic Training Command. The major difference inherent in the experimental grading forms tried out in this study is that they are used to record performance in terms of exactly what a student did during a flight. In the currently used Navy ATJ grading forms performance is rated in terms of judgments of quality of overall performance made in terms of a hypothetical "average" performance.

Both types of grading forms were evaluated utilizing experienced Naval aviators undergoing instructor training as students, and highly experienced instructors from the Instructor Basic Training Unit (Flight). Evaluation in a system such as this enables the investigators to make the statement that unless the forms prove worthwhile in this situation under more or less ideal conditions, there is little hope for their success under less desirable, common operating conditions in the training squadrons. It must also be noted that the success of grading procedures evaluated in this manner does not necessarily preclude their ultimate success in grading Naval Aviation Cadets.

The major conclusions of this study indicate that grading of the experimental type is equally as reliable and valid as the Navy ATJ grading

forms, in terms of consistency and ability to predict future performance. Further, it was conclusively shown that, for the group under observation, the experimental forms offered far less opportunities for distortion of grades through 'halo' effects based on unrelated personality characteristics and knowledge of past performance. The diagnostic value of this experimental grading is also noted.

For certain types of flight performance, it was concluded that the experimental type of grading was superior to the currently used Navy ATJ methods for predictions of short-range future performance. It was further found that the grades on flight proficiencies are somewhat improved by the addition of a subjective judgment of such attributes as "headwork", "mental attitude", "reaction toward flight" and "air discipline". These four attributes are currently a part of the Navy ATJ grading forms, and it is recommended that future grading methods continue to include instructors judgments of these attributes.

It was found that instructor attitudes toward in-flight grading were improved through continued use resulting in better familiarization with the techniques required. It was further found that predictiveness of measurement by the current Navy system of grading was improved during the study. This improvement is felt to be due to an increased awareness on the part of the instructors of the more critical aspects of differential student flight performance.

CHAPTER I

INTRODUCTION

This study represents a further step in the continuing effort on the part of the Navy to develop accurate and meaningful flight grading procedures. Earlier studies in this area are reviewed in Chapter II, following.

Special problems have been encountered in these studies which make difficult the adequate interpretation of results. One of these problems concerns the day-to-day variability in performance of Naval Aviation Cadets. Another problem is the difficulty of properly orienting and indoctrinating flight instructors in the principles and every-day use of objective grading techniques. Still further problems result when recommendations are made regarding the implementation of new grading methods which are based on the results of experimental evaluations.

The variability of daily performance seriously affects the statistical evaluation of grading methods. For example, the determination of reliability is dependent upon the measurement of two identical performances, or the measurement of one performance by two observers. Since flight grading in training aircraft is necessarily limited to one observer, the assessment of reliability must be based on the consistency with which a grading method measures student performance on two, or more, different flights. If a student's performance on two consecutive flights is consistent, his grades on these flights should be highly correlated. Any lack of relationship is then contributable to the unreliability of either the grading

form, or the individual rater. However, since student flight performance is not consistent from day-to-day, the problem is encountered of attempting to reliably measure an unreliable performance.

Criteria for the validation of grading procedures must be based on the prediction of success from some earlier measure. Again the day-to-day variability of student performance causes such predictions to make it impossible to select the particular flight which is most representative of the student's overall ability.

Since it is impossible to isolate the unreliabilities of the student and of the grading method without first developing a highly reliable grading form, it was felt that this study should concentrate on the development of such a form under conditions where a minimum of variability in performance existed. To do this, a group of experienced Naval Aviators in training was selected, on the assumption that their performance on simple flight maneuvers would be relatively stable.

The difficulties involved in the orientation and indoctrination of flight instructors was also felt to be minimized in the Instructor Basic Training Unit, as the Staff Instructors here are a highly select group of instructors experienced in the training of Naval Aviation Cadets.

In conducting this study under somewhat ideal conditions, it must be recognized at the outset that the interpretations of results must be viewed in light of several limitations. First, there is a considerable difference

in using a grading method in a small group, where individual attention can be given to raters, and the implementation of a large scale grading operation in a training unit. Second, the differences between the experienced Naval Aviator and the Cadet may be so great as to make impossible similar grading processes for the two groups. Close attention was paid to the differences between experimental and operating conditions in this study, and the results are interpreted in terms of these differences.

The specific areas which are investigated in this research involve the problems of in-flight vs. post-flight grading, reports of actual performance vs. subjective judgments of qualitative proficiency, and instructor acceptance of newly-designed grading techniques. Specific questions to which answers are sought include the following:

1. By what standards are students currently measured and how reliable is this measurement?

2. How well can check-flight performance be predicted from daily performance within a stage of training?

3. Can a grading system feasibly be developed which minimizes the influence of non-related effects of student upon instructor?

4. Can a grading system be devised which would be equally as efficient with new instructors as with experienced ones?

5. What must be done in order to implement a new method of grading?

The answers to these, and other questions, are made on the basis of experimental conditions, and must be modified in terms of training conditions. However, since the conditions under which this study is conducted are considerably more favorable, it can be stated that, unless a grading system is found to be reliable, valid and useable here, there is little hope for its success under less favorable conditions in the cadet training squadrons.

CHAPTER II

HISTORICAL BACKGROUND

Uses of Flight Grading Techniques

The effectiveness of any training program is related to the accuracy with which measures of proficiency of its students reflect their actual ability. Measures of flight performance, during both training and operational tasks, make up the foundation upon which administrative training decisions are made. Without accurate measures of proficiency, little could be said about student's readiness to attempt more difficult phases of training, or about the feasibility of investing extra time, effort and money in borderline cases. If such measurement is accurate, a sound basis is provided for making such decisions concerning advancement.

A further need for improved measures of proficiency, particularly in highly complex tasks such as flying modern aircraft is in the area of diagnostic evaluation. Here, reference is made to the grading of details of performance in such a manner that a student's record will reflect his specific strengths and weaknesses. If these measures are diagnostic in nature, their results can be used to gear further training to the specific needs of the individual student.

Of importance also is the use of proficiency measures taken during training and in ultimate assignments. These measures are necessary for determining the value of differing types of training procedures, use of

special training devices, and research on the usefulness of pilot candidate selection methods.

This project is a part of the continuing effort on the part of the Naval Air Training Command to provide usable methods of measuring flying performance accurately. It is hoped that measurement can be improved so that it may serve as a sound basis both for administrative decisions and for future research in the improvement of selection and training.

The Need for Improvement in Flight Grades

For many years subjective ratings of flying ability have been used as the basis for grades in Naval Air Training on all instructional flights. These subjective grades, along with the written comments of the flight instructor, make up the largest portion of available information on which decisions are made regarding attrition or retention of borderline students. Research in the development of new selection tests and evaluation of training programs and training devices must depend for criteria on the numerical scores derived from these subjective grades and on the fact that a student completed training or attrited from the program.

In connection with a large scale research project designed to provide improved selection tests for Naval Air Training, The Psychological Corporation undertook, at the request of the Bureau of Medicine and Surgery, a project to determine the reliability and validity of the above-mention-

ed subjective grades. That study (1) published in May 1949, involved a statistical analysis of data from the flight training jackets of 357 Naval Aviation Cadets in Basic Training. In the course of this analysis, it became clear that the subjective rating scales used to determine flight grades were not highly reliable. They did not furnish satisfactory predictions of later proficiency, nor adequate criteria for the evaluation of selection tests and training experiments.

These conclusions concerning grades in Basic Training were confirmed in a similar study of Naval Air Advanced Training grades (8) published in 1950. Additional evidence from many sources supports the general conclusion that flight grades based exclusively on subjective ratings are inadequate for predicting success in the training program.

Indications that Improvements are Possible

In addition to pointing out deficiencies, research on the problems of flight grading has suggested many methods of improvement. One of the most promising techniques is the standardized flight and objective record. This method was developed early in World War II by the committee of Aviation Psychology, National Research Council (11), and has received a large share of the attention of aviation research personnel during the past few years.

The most successful application of this method was Gordon's (3) development of a standard flight check for rating the airline transport pilot.

The reliability of this check flight, as determined in a tryout with experienced pilots, is the highest ever reported for two successive check rides graded by two different check pilots, being .58 in one study and .76 in a later one (3). In a further revision of this check flight, a ride-ride reliability of .71 was obtained for two successive check rides graded by two different raters using airline pilots as subjects (10).

Description of Parent Study

The U. S. Naval School of Aviation Medicine and The Psychological Corporation conducted a joint project in 1951 to develop and evaluate objective, in-flight grading methods for two stages of Naval Air Training[1] (12). The plan of that study followed the plan utilized by Gordon (3), with certain modifications necessitated by the differences in the task performed. Modifications were made also in light of the findings of previous Navy flight grading research. The general approach involved:

1. Concentration on standardized check flights as the primary measure of proficiency.
2. Obtaining an itemized, objective record based on what the student actually did during the flight.
3. In-flight marking of the performance as it occurred or as soon thereafter as possible.
4. Insuring a clear definition of the maneuvers to be performed and the manner in which they were to be graded.

The major emphasis in this study was placed on determining the reliability

[1] The two stages selected were pre-solo primary and basic instrument. Pre-solo primary was chosen because most flight failures occur there, making any improvement in accuracy of grading highly desirable. The basic instrument stage was selected since instrument flying appeared to lend itself easily to objective grading.

of the check flight grade on two successive, identical flights administered by two different flight instructors. In addition to this, internal reliabilities of the forms were determined, and instructor comments were collected regarding the utility of such a grading system for routine use.

The results of this study revealed that the grading measures developed and tried out did not prove superior to the currently used subjective measures. The forms did, however, possess fairly high internal reliabilities. Even if the forms had been revealed to have high ride-ride reliability, no recommendation would have been made for their routine use in the training situation without considerable simplification. The reason for this was that the majority of instructors reported that they believed the complexity of the forms caused them to be dangerous for daily in-flight grading.

Planning the Present Study

World War II studies in the Army Air Forces revealed that attempts to improve ride-ride reliability often failed because of erratic day-to-day fluctuations in performance, rather than because of measurement errors (9, p.361). The Army Air Force and the Navy, in their attempts to improve reliability of flight grades utilized trainees in their respective programs as subjects for their studies. Gordon (3) and Nagay (10) in the Airline Transport Rating studies, where high ride-ride reliabilities were reported, measured the performance of experienced pilots. Wilcoxon and Johnson (12) state, concerning the possible reasons for unreliability of their grading form:

> "If we may suppose that the grading forms used in this investigation are representative of those used elsewhere, it appears that the most likely reason for low ride-ride reliability lies in the variability of student performance from one ride to the next."

On the basis of these results, the present project was designed to utilize Instructors Under Training from the Instructor Basic Training Unit (Flight) as subjects. These pilots are experienced Naval aviators, and therefore, were assumed to be more stable in their day-to-day flying performance than cadets. It was also designed to collect measures of flying ability on each of the instructional flights in a particular stage of training, and to simplify these measures in order to increase their useability. It is fortunate, also, that the training syllabus of this unit allows for sequential measures of proficiency during instructional flights.

CHAPTER III

DESCRIPTION OF THE STUDY

The Instructor Basic Training Unit

This unit serves two major functions in the Naval Air Basic Training Command. First, the unit serves to train Naval aviators in the methods of instructing cadets in the various maneuvers taught in the different stages of basic training. Second, the unit serves as a standardization control of the content of the stages and methods of instruction of these maneuvers. The unit is divided into a ground division and a flight division. The ground division conducts lectures on the principles of instruction, practice in the elements of speech and indoctrination and orientation to the policies of the Naval Air Basic Training Command. Its program is approximately two weeks in length. The flight division conducts training and practice in the maneuvers which must be taught in the air, and requires approximately six weeks.

Instructors Under Training in the Primary Stage of the Instructor Basic Training Unit (Flight) perform the identical maneuvers that Naval Aviation Cadets will be required to perform in the Primary syllabus at Whiting and Corry Fields. After successfully completing this Primary syllabus, including two check flights, one on high work and one on low work, the Instructors Under Training advance to a second stage which involves the

learning and practice of instructional techniques of these maneuvers. This alternating plan of performing, then teaching, continues throughout the several stages of work.

The Primary stage was selected for evaluation for two reasons. First, it is closely related in content to the Pre-solo Primary Stage of cadet training. Second, it offers an opportunity to observe Instructors Under Training in the early stages of learning and development. The learning situation here is actually one of re-learning the tasks in which these Naval Aviators were trained earlier in their careers. In these stages, concentration is on flying skills, rather than on the learning of instructional techniques. The latter process is inherent in most other stages of the Instructor Basic Training Unit, and its measurement is not evaluated herein.

The relationship of instructor to student offers an important contrast in this research. The instructors attached to the Instructor Basic Training Unit have been selected from the various outlying units on the basis of experience and proficiency. The Instructors Under Training, which we shall subsequently refer to as 'students' throughout this report, are experienced Naval aviators reporting to the Training Command and thus, are different in many respects from the typical Naval Aviation Cadet learning to fly for the first time.

In this experimental tryout of a new grading system, therefore, more or less ideal conditions of instructors and students exists. The observations

and recordings of grades are made by a select group of experienced flight instructors, and the students who are being observed are designated pilots who should be less variable in their day-to-day performance than cadets. The limitations due to these differences between the experimental and the real training situation are considered below in the conclusions drawn from the results of this study.

The Primary Syllabus

The maneuvers performed in the Primary stage are divided into two somewhat contrasting types of work. The first two flights, A1 and A3, are performed at low altitude and consist of practice in landings, takeoffs and field procedures required in order that the student may be considered 'safe for solo'. The next three flights, A4, A5 and A6, are performed at a high altitude and include instruction and practice in maneuvers such as spins, stalls, spirals, etc., in preparation for the A7 high work check flight. The next three syllabus flights, A8, A9 and A12, deal with the low work contained in the A13 check flight. This includes such maneuvers as touch-and-go landings, small field procedures, standard field entries and low altitude emergencies, among others.

In correlating the grades made on the low work and high work sections of the Primary syllabus, fairly low relationships were found, indicating that the material covered on these two parts is somewhat independent. Statistical analyses in this study are therefore made separately on the two parts of the syllabus.

The ATJ Flight Grading Forms

The standard method of flight grading in the Navy is a system based on instructor judgment of student performance in terms of individual concepts of a hypothetical average student. The ATJ grading form appearing in Appendix A is representative of most currently used Navy grading forms. It contains a listing of the maneuvers performed in the Primary stage, as well as a listing of four general attributes. These attributes are "headwork", "mental attitude", "reaction toward flight" and "air discipline".

On each of the indices of flying skill and attributes included in the ATJ form, the instructor may rate his student in terms of deviations from "average" in four possible categories. These categories are "unsatisfactory", "below average", "average" or "above average". The large majority of all marks assigned are in the "average" category, with deviations upward or downward checked when appropriate. No set number of items to be graded is prescribed for a particular flight, rather, the instructor checks only those items which pertain to the flight. The ATJ grading form is filled out on the ground at some time after the flight has been completed. In most units of the Training Command, it has been observed that these forms are generally filled out on the same day as the flight, or within two to three days following the flight.

These forms are scored by totaling the number of marks assigned in each of the four categories. No other numerical assessment of flight per-

formance is made for an individual flight. A cumulative total is kept of these marks and a student's grade for a specific stage as well as his final grade in the Basic Training Command is based on the sum total of the number of marks in each of these four categories.

In this study, the ATJ grades which are used for comparative purposes have been broken down into two separate grades. The first is based on the ratings of flight performance contained in the first part of the form and the second is based on the attributes assessed by the last four items of the form. It is felt that certain differences are inherent in these two factors and, therefore, grades should be computed separately.

The Experimental Design

In this study, experimental grading forms were developed covering many aspects of the maneuvers of the Primary syllabus. These forms were developed through cooperation with a Pilot Advisory Board. This board consisted of three staff instructors selected from the Instructor Basic Training Unit (Flight). These grading forms included those aspects of performance which were felt by the board and the investigators to be measurable and pertinent to flight proficiency measurement. The experimental grading forms were filled out in the air, followed by routine grading with the Navy ATJ forms at some time following the flight.

For purposes of comparison, two groups were selected for study. The first was a control group, consisting of one hundred and sixty students in the

Instructor Basic Training Unit (Flight) from April 12, 1953 through August 10, 1953. Records were examined and analyses made of the grades of these students on the standard Navy ATJ grading form. The second, or experimental group, consisted of the population of sixty-seven students trained during the period from December 1, 1953 through February 19, 1954. The experimental group was graded by both the Navy ATJ form and by the newly developed experimental grading forms.

The staff instructors participating in this study were selected on a volunteer basis and were thoroughly indoctrinated in the use of the experimental grading forms prior to the flights. Analyses were made of these grading data and comparisons were made which are reported in Chapter V, following. At the close of the study, questionnaires were administered to participating instructors in order to obtain information pertinent to the evaluation.

CHAPTER IV

PROCEDURES INVOLVED

Development of the Experimental Grading Forms

To assist in determining the maneuvers to be graded in this study, a three-man panel was selected. The members of this board were chosen from volunteers on the basis of being 'typical' of the majority of the Staff Instructors in the unit. The investigators held numerous meetings with this board and examined every flight contained in the Primary syllabus. Determinations were made of those aspects of student performance which were both measurable and pertinent from the standpoint of flight proficiency evaluation.

It was found that many critical aspects of flight do not lend themselves to purely objective grading; that is, grading based on actual measurement rather than on observer judgment.[1] For these aspects, an attempt was made to construct the forms in such a way that instructors reported what a student did, rather than a judgment of the quality of such performance. This often required a response such as "proper" or "improper" regarding the performance of a specific task. Although such items require judgments to be made by the instructors, these judgments differ from the purely subjective to reportings based on performance with respect to a standard. The judgments made on the Navy ATJ forms are made with

---

[1] An example of pure objective measurement is the recording of an instrument or dial reading, while subjective measurement is exemplified by an assessment of a performance by an observer in terms of "good" or "poor".

respect to an individual instructor's concept of the "average" performance. In cases where more objectivity was possible, items were constructed in the form of recordings of dial readings.

The original experimental grading forms drawn up as a result of these conferences were filled out in the air on several flights during initial tryouts by the members of the advisory board. After these trials were completed, further conferences were held in which items were revised and necessary additions and deletions made. A second revision was then tried out in a similar manner, and in some cases several more trials were necessary in order to refine the forms for each maneuver so that they could be presented to the unit for use. The final revisions of each of the forms appear in Appendix B and were gathered into booklets of approximately four or five pages, each booklet representing a flight.

The forms were designed with a cardboard backing which slipped into a specially designed knee pad. This arrangement made it possible for the flight instructor to turn the pages one at a time and securely fasten them in back with a relatively simple motion. The knee pads constructed for use in this study were similar to those developed by Wilcoxon and Johnson (12), and are fully described in that work.

Administration of Flight Grading

A group of ten volunteer staff instructors were selected to perform the grading of students in the experimental group. These instructors were

carefully indoctrinated in the exact procedures required for filling out the experimental grading forms. Care was taken in cases where judgments of proper or improper performances were to be made to see that the group as a whole was standardized with respect to what should constitute a correct, or an incorrect, procedure.

As the number of students in the experimental group increased it became necessary to add to this nucleus of ten original instructors until a total of twenty-eight instructors were regularly being used toward the latter part of the study. As each instructor was added to the group, he was individually indoctrinated in the use of each of the grading forms. The experimental grading forms were distributed as a part of the checking out routine before each flight and were returned immediately following the flight.

Comparison of Number of Items Included in the Two Grading Forms

The number of items graded on each flight of the Primary syllabus by the experimental grading forms is presented in Table I and is therein contrasted with the number of items graded on the ATJ forms.

As no specific number of items is required in completing the Navy ATJ forms, the number of items indicated in Table I represents the modal number filled out for each of the flights. While there was considerable variance with respect to the individual number of items checked by different instructors, all flights required that the four attribute items at the end of the form be filled out.

It is noted that the experimental grading forms contain a considerably larger number of items on which the student is rated. An example of the

TABLE I

NUMBER OF ITEMS GRADED ON EACH SYLLABUS FLIGHT BY ATJ AND EXPERIMENTAL FORMS

FLIGHT	DESCRIPTION	NUMBER OF ITEMS	
		ATJ	Experimental
A1	Safe for solo check	9	39
A3	Intoductory low work in precision approaches and no-flap landings	10	95
A4	High work - stalls, steep turns and precision spins	9	58
A5	Stalls, slow flight, spirals and high altitude emergencies	7	32
A6	Review of high work	9	96
A7X	High work check flight	14	96
A8	Low work - small field procedures and low altitude emergencies	6	57
A9	Touch and go landings, precision approaches and cross-wind landings	9	85
A12	Review of low work	12	116
A13X	Low work check flight	14	116

grading of stalls should suffice for explanation of this difference. On the Navy ATJ form, one item is provided for the overall assessment of performance on all stall maneuvers. On the experimental forms, each stall performed is graded separately in terms of performance on six different aspects. Thus, on the A6 flight, containing nine different types of stalls, one item is graded on the Navy ATJ form, and fifty-four on the experimental form.

Scoring of the Experimental Grading Forms

The items in the experimental grading forms differed among themselves with respect to possibilities of scoring. Some of the items lent themselves to a dichotomy of "yes" and "no", "proper" and "improper", etc. Others contained three or four choices of varying degrees of correctness of performance. Still others included dial readings in the form of marks on a scale with the center point as the most desirable reading for a particular aspect of a maneuver.

Danneskiold (2), in an earlier flight grading study, found that items of this type could be graded either according to a multi-point scale of quality of performance or they could be dichotomized into a correct or an incorrect response based on group performance. In that study, resultant scores on total forms were intercorrelated by coefficients ranging from .97 to .99. In view of this high degree of similarity, each of the items of the experimental grading form was broken into its various possible answers and such answers were tallied for the group under study.

An examination of these group responses was then made to determine a cut-off point which would separate correct from incorrect responses. Dichotomies were assigned to every item so that, as nearly as possible, a grade of 'correct' was attained by approximately fifty per cent of the group. In this way, the item difficulty for each item and for the entire form was adjusted to be approximately .50.

The initial score obtained for each student on each flight was a simple percentage score representing the total number of items scored correctly divided by the total number of items graded. This percentage figure was necessary due to the fact that in some cases certain parts of a maneuver were omitted. An example of these omissions occurred when a touch-and-go landing was graded up to the point of touchdown and at that point the student would be given a wave-off, either due to his own error or due to existing conditions in the traffic pattern. In such cases, approximately 60 per cent of the items for that particular landing were graded with the remainder being left blank. Another example appears on the high work check flight where nine stalls are normally performed. In a few cases, an instructor may require only seven or eight of these nine stalls to be performed, thus causing a slight difference in the total number of items answered for the flight as a whole. In no case, however, was a flight considered graded for purposes of this study if more than 25 per cent of the total possible items were omitted for any reason.

Alternate Scoring of Experimental Grading Forms

Many comments were made by instructors during the study that certain aspects of a maneuver were distinctly more important than others in determining student ability. Comments were also heard that certain maneuvers within a flight were more critical than others. For this reason it was decided to attempt to develop a meaningful method of weighting various items and maneuvers to obtain the score which would best represent the quality of student performance.

To accomplish this, questionnaires were distributed near the close of the study which contained a listing of every item of every maneuver, every maneuver of every flight and every flight of the syllabus. For these listings, the instructors were asked to select the most important and the least important of these items and maneuvers. Instructions were given that such judgments of importance were to be based upon factors which can be used to differentiate between good and poor students. The questionnaires were administered under group conditions in a classroom of the training unit, and names were not required on the sheets.

Tallies were made of the responses and corresponding weights were assigned to the items and maneuvers. In many cases, high agreement was found among the twenty-eight experimental instructors participating in the survey. Instances were found where 90 per cent or more of the group felt that a particular item was most important and another item of the same maneuver was considered least important by a similar percentage. In some cases,

items and maneuvers appeared to be considered equally important by the group.

The experimental forms were then scored in accordance with these weights. These weighted scores were found to correlate .94 with the simple raw scores previously obtained. This correlation was computed by the Pearson product moment method for all flights for the entire group. This high degree of similarity between the two scores indicates that there is little advantage in weighting scores on performance of this type. However, since it is reasonable to assume that any differences between the two scores would undoubtedly favor the weighted score as a more representative indication of quality of student performance, it was used throughout the statistical analyses described below.

Questionnaires for Further Analysis of Instructor Opinion

In order to assess the reactions of flight instructors to the experimental grading forms used in this study, the above questionnaires also included provisions for unstructured comments, as well as for answers to the following four questions:

1. Do you feel that these grading forms would be a detriment to safety in the air

in IBTU?		at Whiting Field?[1]	
Yes ☐	No ☐	Yes ☐	No ☐

[1] Whiting Field is the location of Basic Training Units 1(N) and 1(S), in which the majority of all Naval Aviation Cadets receive their Primary Pre-Solo training.

2. Do you feel that these grading forms are superior to the UBAA forms for

	Yes	No
de-briefing the student?	☐	☐
board reviews of jackets?	☐	☐
permanent proficiency records?	☐	☐

3. Do you feel that these grading forms are most valuable for:

daily instructional hops?	☐
check flights?	☐
both?	☐
neither?	☐

4. Do you feel that the grading forms could be shortened and still give an accurate account of what the student did?

Yes ☐ No ☐

Responses to these items were tallied and a classification was made of the unstructured comments, all of which appear in Chapter V, below.

Scoring of the ATJ Forms

The Navy ATJ grading forms on which students in both the experimental and control groups were graded were scored in terms of two part scores and a total score. The first of these part scores was an index of flight proficiency indicated by a weighted sum of the responses to each of the air-work items included in the first part of the ATJ form. This score was obtained by weighting the total number of responses in each of the four

categories. Responses in the unsatisfactory category were given weights of 0, those in the below average category weights of 1, those in the average category weights of 2 and those in the above average category weights of 3.

A grand total was then obtained for this part of the form, which, in turn, was divided by the number of items, giving an average score for flight proficiency. This averaging was necessary again on this grading form, as in the case of the experimental form, due to the fact that different numbers of items were answered for the same flight by different instructors.

The second part score of the ATJ grading form consisted of a similar totaling and averaging process for the four attributes. The relationships between these attribute scores and the flight scores for the control and experimental groups are indicated in Table II. The correlations presented

TABLE II

CORRELATIONS BETWEEN PART SCORES ON THE ATJ GRADING FORMS FOR CONTROL AND EXPERIMENTAL GROUPS

GROUP	HIGH WORK		LOW WORK	
	DAILY	CHECK	DAILY	CHECK
CONTROL	.34	.44	.17	.31
EXPERIMENTAL	.33	.36	.12	.34

are Pearson product moment coefficients, and were computed for daily flights between the two part scores on a composite total daily score for low and high work individually. These scores were obtained by converting each of the flight scores to a standard score for that particular flight, and adding these standard scores.

The relationships are seen to be fairly low between the two measures, indicating that they are somewhat independent. It was therefore decided to evaluate the ATJ grade by means of flight and attribute grades separately. A third score for the ATJ grading forms was obtained by averaging all ratings given on the flight for an average total grade. This total score corresponds with current scoring practices in the Training Command.

CHAPTER V

RESEARCH FINDINGS

Relationships Between Experimental and ATJ Grades

In a study of this type where an experimental method of grading is being compared with an existing method, it is important to note at the outset the relationships existing between the two methods of grading. If an extremely high correlation is found between the grades assigned by the two different grading forms then it can be assumed that little improvement is possible since the two forms are measuring essentially the same thing. Pearson product moment coefficients of correlation were computed between scores obtained on the ATJ forms and on the experimental forms for daily flights and also for the two check flights. These correlations are presented in Table III.

TABLE III

RELATIONSHIPS BETWEEN TOTAL SCORES ON ATJ AND EXPERIMENTAL GRADES FOR DAILY AND CHECK FLIGHTS (N=67)

DAILY GRADES		CHECK FLIGHT GRADES	
High Work	.43	High Work Check Flight	.62
Low Work	.59	Low Work Check Flight	.69

These relationships between the two methods of grading appear to be slightly

higher for a single check flight than for a combination of daily flights. This is explained by the fact that student fluctuations in performance from one flight to the next are averaged out when several daily scores are combined. The resultant correlations between two methods of grading are therefore lowered. The correlations, in general, indicate some differences in performance measurement are obtained on the experimental grading form. These differences are not as distinct as they appear, since the individual reliabilities were found to be .63 and .71 for the ATJ and experimental forms, respectively.

The above relationships, having been computed for total scores, include the common element of the attribute score in each, and are representative of the grades which would ultimately be used in the Basic Training Command (see page 35). Considering grades made on flight performance only, somewhat lower relationships are presented in Table IV. It appears that measurements of considerably different flight performances are obtained by the two methods.

TABLE IV

RELATIONSHIPS BETWEEN FLIGHT SCORES ON ATJ AND EXPERIMENTAL GRADES FOR DAILY AND CHECK FLIGHTS (N=67)

DAILY GRADES		CHECK FLIGHT GRADES	
High Work	.32	High Work Check Flight	.48
Low Work	.40	Low Work Check Flight	.46

Differences in Scores Between "Up" Flights and "Down" Flights

In the Naval Air Basic Training Command, each instructional flight and check flight is assigned one overall judgment of passing or failing. This judgment is made by the individual flight instructor, and is based on his general assessment of the student's performance. This judgment may be made either in terms of performance of required maneuvers, or in terms of attitudes and reactions toward instruction. This general assessment results in the flight being labeled either "up" or "down".

In the two groups studied, only 3.7 per cent of the flights in the control group and 4.0 per cent of the flights in the experimental group were judged as "downs". Means and standard deviations were calculated for all scores on both grading methods for flights judged as "up" and flights judged as "down", in order to determine whether or not the grades of these flights differed significantly by each of the methods of grading.

Each of the differences between mean scores on "up" flights and on "down" flights in Table V are statistically significant at better than the 1 per cent level of confidence throughout all grades of both samples. The performance grades assigned by both methods of grading, therefore, produce significantly lower scores for those flights judged as "down" than for those flights judged as "up". It is interesting to note that on the ATJ grades the difference between an "up" flight and a "down" flight is considerably greater on flying skills than on attributes.

Reliability of Measurement

Of primary concern in the evaluation of a grading system are its reliability and its ability to predict future performance. The concept of

TABLE V

MEANS AND STANDARD DEVIATIONS OF SCORES ON PASSING AND FAILING FLIGHTS

	CONTROL GROUP (N = 160)		EXPERIMENTAL GROUP (N = 67)			
	ATJ GRADES		ATJ GRADES		EXPERIMENTAL GRADES	
	M	SD	M	SD	M	SD
FLIGHT SCORE:						
"Up"	196.01	27.14	197.72	26.16	70.56	16.96
"Down"	157.53	27.20	147.35	28.44	50.29	13.85
ATTRIBUTE SCORE:						
"Up"	218.70	20.92	221.59	21.11		
"Down"	195.42	16.37	192.31	16.62		
TOTAL SCORE:						
"Up"	204.27	20.38	206.61	20.16	70.76	13.85
"Down"	157.88	22.15	160.46	25.30	51.33	16.49

reliability embraces the idea that a grading system, in order to be worthwhile, must be accurate and consistent.

There are three general methods by which the reliability of a test is assessed. One of these procedures involves the administration of a test to

a group followed by a later administration of the same test to the same group. In this test-retest situation, there should be a high relationship between scores made by individuals at each of the testing periods. Another index of reliability can be obtained by administering similar forms of the same test to a group. Here, the relationships between the scores on the two similar tests should be high.

The most commonly used method of determining reliability involves the split-half principle. Here, a test is split into two chance halves, and the relationship of these parts is computed. These halves are often odd items and even items, or the chance halves may be selected by other random, or stratified random methods.

In determining the reliability of flight grading methods the measure of ride-ride reliability has often been used. As discussed above, this ride-ride reliability figure may often be low due to the actual variability of student aviators in their performance from one flight to the next. If these fluctuations are great, the ride-ride reliability of a grading form, no matter how accurate and consistent it is, will necessarily be low.

An attempt was made in this study to overcome such fluctuations by using experienced Naval aviators as students, and by averaging several daily performances. The split-half reliability determined here was one of a composite of all grades in the Primary syllabus.

In correlating the odd-numbered flights against the even-numbered flights,

the theory is considered that if the grading forms were entirely unreliable, the mark assigned on these forms would be no better than chance grading, and the correlation between odd flights and even flights in the syllabus would be zero.

Each of the scores made by both methods of grading in the experimental group and by the ATJ method in the control group was converted to standard scores and combined so that each student received a total grade for the odd-numbered flights and a total grade for the even-numbered flights.

The correlations between the odd-and even-numbered flights, since they represent relationships between one half of the syllabus and the other half are boosted by means of the Spearman-Brown Prophecy Formula (4:275), and are presented in Table VI.

TABLE VI

SPLIT-HALF RELIABILITIES OF ATJ AND EXPERIMENTAL GRADING FORMS FOR THE PRIMARY SYLLABUS

	CONTROL GROUP	EXPERIMENTAL GROUP	
	ATJ GRADES	ATJ GRADES	EXPERIMENTAL GRADES
Flight	.63	.61	.70
Attribute	.42	.59	
Total	.57	.63	.71

An examination of the split-half reliabilities in Table VI reveal that the flight grades obtained on the experimental forms are slightly higher than those obtained on the ATJ grading forms. This difference, .70 over .61, is not a statistically significant one, as determined by a conservative test (5:175ff). The reliabilities of both forms, however, are considerably higher than those previously found on a ride-ride basis.

In considering the grading of flight attributes, it is noted that the attribute scores on the ATJ grading forms for the control group were significantly less reliable than the flight grades for this group. In the experimental group, however, this was not the case. Since the grading of attributes appeared to be equally as reliable as the grading of flight performance in the experimental group, and since no provision was made in the experimental grading forms for assessment of these attributes, it was decided to combine the experimental flight scores with the ATJ attribute scores in the hopes of obtaining a more stable index of student performance. This combination was also done at the request of the flight instructors participating in the study. Many of these instructors expressed the desire to provide for grading of these attributes on the experimental grading forms. The reliability of the experimental grading form was not appreciably changed by the addition of these flight attributes.

Predictiveness of Measurement

In order to be useful, a grading form must be able to predict future performance. The greatest need for prediction in flight training is in determining the long range future success or failure of a student pilot.

Some indication of this predictiveness can be obtained by studying the short range validity of a grading form.

One of the primary interests of this study was to determine if a grading system could be developed which would predict student performance on a check flight from his performance on preceeding daily flights. In order to accomplish this, the three daily flights of each part of the syllabus were combined by adding standard scores to produce a daily score, which in turn was correlated with the check flight score. The A4, A5 and A6 flights composed the entire high work syllabus which is evaluated in the A7 check flight. Low work performance is taught on the A8, A9 and A12 flights which prepare the student for his A13 check. The correlations thus obtained between daily flights and check flights are reported in Tables VII and VIII, and show that the experimental grading forms do not predict the A7 check flight to as great an extent as does the ATJ method of grading. The differences are not significant, however. For the prediction of the A13 check flight, which is composed of low work including landings and takeoffs, the experimental grading form appears to be somewhat superior.

It is especially interesting to note that in both predictions the validity of the experimental form was increased by adding the attribute score to the scores obtained on actual flight performance. The prediction of the A13 check from the composite score of the A8 through A12 flights of .62 is a very high one, considering the reliability of the form, and is significantly higher than any obtained by ATJ grades.

The predictions of ATJ grades on both check flights from their respective daily grades appear to be slightly higher in the experimental group than in the control group.

TABLE VII

PREDICTION OF A7 HIGH WORK CHECK FROM
TOTAL DAILY LOW WORK SCORES

	CONTROL GROUP	EXPERIMENTAL GROUP	
	ATJ GRADES	ATJ GRADES	EXPERIMENTAL GRADES
Flight	.46	.50	.34
Attribute	.22	.26	
Total	.41	.43	.39

TABLE VIII

PREDICTION OF A13 LOW WORK CHECK FROM
TOTAL DAILY HIGH WORK GRADES

	CONTROL GROUP	EXPERIMENTAL GROUP	
	ATJ GRADES	ATJ GRADES	EXPERIMENTAL GRADES
Flight	.28	.34	.45
Attribute	.00	.15	
Total	.19	.28	.62

A Halo Analysis

The effects of intangible and often unrelated characteristics of a student upon an instructor are referred to as halo effects. It is not uncommon for a well-groomed and eager appearing student with a good personality to create such a favorable impression upon his instructor that his actual performance is rated higher than is objectively warranted. Similarly, the student who makes an unfavorable impression may be unjustly rated down in his actual performance. Also, the quality of a previous performance may often influence an instructor's judgment of a later performance.

In Navy flight grading it is well known that a student's past performance plays an important part in many cases in the grade which he is assigned on a particular flight. Opportunities for this type of distortion of grades are greatest where evaluations are based on a subjective judgment of general quality of performance.

It must be noted, of course, that student performance may, in fact, be affected by the presence or absense of a familiar instructor. This effect of instructor on student is independent of the grading methods employed, and such methods should reflect actual performance, rather than performance modified by the effect of student upon instructor.

In order to evaluate the comparative halo effect of the ATJ grading forms and of the experimental forms analyzed in this study, an analysis was undertaken to compare the grades made on consecutive flights given by the same instructor and to compare the grades made on consecutive flights

given by different instructors. In the Instructor Basic Training Unit, the students are usually assigned different staff instructors for each flight. However, instances were noted where two consecutive flights were given a student by the same instructor. Since the instructors are assigned more or less by chance and there is no systematic variable operating, it was felt that comparisons could be made on this basis, according to the following plan:

1. Of the 1648 flights recorded for the control group, 53 pairs of consecutive flights were found where a student was graded by the same instructor.

2. Pearson product moment coefficients of correlation were computed between the grades made on the first and the second flights selected in (1), above.

3. Of the 646 flights recorded in the experimental group, 46 pairs of consecutive flights were found to have been given the same student by the same instructor.

4. Pearson product moment coefficients of correlation were computed between the grades made on the first and the second flights selected in (3), above.

5. Pairs of consecutive flights were then selected at random in both groups from among the remaining pairs of consecutive flights given by different instructors.

    a. This random selection was stratified to maintain the same proportion of syllabus numbered flights as was found in (1) and (3) above.

    b. This selection was also stratified with respect to instructor so that one of the instructors of each pair corresponded with that selected in (1) and (3), above.

6. Pearson product moment coefficients of correlation were computed between the first and the second flights of the pairs selected in (5), above.

The resultant coefficients of correlation are presented in Table IX.

An examination of these relationships reveals a distinct contrast in both the control group and the experimental group between the relationships of consecutive flights graded by the same instructor and consecutive flights graded by different instructors on ATJ grades. The total score relationship is .38 in the control group and .42 in the experimental

TABLE IX

CORRELATIONS BETWEEN PAIRS OF FLIGHTS GRADED BY THE SAME INSTRUCTOR AND BY DIFFERENT INSTRUCTORS

	CONTROL GROUP		EXPERIMENTAL GROUP			
	ATJ GRADES		ATJ GRADES		EXPERIMENTAL GRADES	
	Same	Different	Same	Different	Same	Different
FLIGHT	.36	.07 *	.37	.11 *	.45	.36
ATTRIBUTE	.48	.05 *	.41	.04 *		
TOTAL	.38	.01 *	.42	.10 *	.45	.32

* (Difference significant at .01 level)

group between pairs of flights graded by the same instructor. Similarly, a relationship of .45 existed between experimental grades.

The resultant correlation between pairs of consecutive flights given by different instructors shows a startling contrast when graded by the ATJ method in both the control and experimental groups. The total score correlations were .01 and .10. Each of these differs significantly from the

respective correlations obtained under same-instructor conditions. The significance of these differences is at the .01 level. However, when the objective grades made on the same flights were correlated for the pairs given by different instructors a relationship of .36 was found, the difference between .36 and .45 for this size sample is not a significant one. While there appears to be some effect of the same instructor in the grading of the experimental forms it appears to be considerably less than the significant effect noted on the ATJ grades.

Analysis of Instructor Opinion

The questionnaires described in Chapter III, above, were analyzed in several ways and revealed opinions of the flight instructors participating in the study. Regarding the hazards involved in the use of the experimental in-flight grading methods, 15 out of the 28 instructors felt that these grading forms were unsafe for use in the Instructor Basic Training Unit, while 13 felt they were safe. Concerning their possible use at Whiting Field for grading Naval cadets in the Primary stage, 26 out of 28 felt they were unsafe.

On the item referring to the possible superiority of the experimental grading forms to the ATJ forms for specific purposes, out of 28 instructors, the following opinions were found:

a. 21 instructors felt that the experimental grading forms were superior to the ATJ grading forms for debriefing the student immediately following the flight.

b. 22 instructors felt that the experimental grading forms were superior for board reviews of student performance, for determining whether or not a student should be allowed to continue in the program.

c. 17 instructors felt that the experimental forms were superior for permanent proficiency records.

Numerous comments were also noted by the authors of the usefulness of these forms in diagnostic reviews of student deficiencies.

When asked if the experimental grading forms were most valuable for daily instructional flights, for check flights, or for both, the following break-down of responses was obtained:

Daily instructional flights	1
Check Flights	18
Both	7
Neither	2
Total Responses	28

There appeared to be a definite feeling that these forms were most valuable for check flights. One reason for this is that on a check flight no instruction takes place. The student is merely observed and rated on his ability to perform the required maneuvers. In contrast to this, instructional flights require that the majority of flight time is devoted to actual instruction of the student. Comments were noted by the authors to the effect that filling out the grading form interferred with such instruction.

On the question of whether or not the experimental grading forms could be shortened and still give an accurate account of what the student did, the group poll was divided fifty-fifty on the "yes" and "no" responses.

A hypothesis was considered that instructor feeling toward, and opinions regarding, these grading forms might be effected by the amount of experience

and the length of time with which they used the forms. In order to evaluate the possible relationship between responses to the questionnaire items and experience in use of the forms, the group of instructors was dichotomized, both in terms of number of forms filled out, and in terms of length of time served in the experimental study. Instructors ranged from six to forty-one with respect to number of experimental grading forms filled out in this study, while a range of two weeks to twelve weeks was found in terms of participation in the study.

The resultant dichotomies were evaluated according to methods set forth by Kendall (6:35,44). These relationships are summarized in Tables X and XI and indicate certain significant trends.

TABLE X

RELATIONSHIPS BETWEEN INSTRUCTOR OPINION AND NUMBER OF FORMS GRADED

RELATIONSHIPS TESTED	R*	S/varS**
Number of flights graded vs. safety opinion	.20	.66
Number of flights graded vs. superiority over ATJ	.52	2.21 (.05)
Number of flights graded vs. value for check flights	.54	1.98 (.05)
Number of flights graded vs. shortening of forms	.62	2.22 (.05)

* Estimate of Pearson product moment coefficient of correlation from tau value.

** Positive products minus negative products of a two-fold table divided by the respective variance (6:44).

On the question of safety of in-flight grading, no significant relationships were noted, nor was a significant relationship noted between opinions regarding shortening of the forms and length of time exposed to this method of grading. On each of the other questions, however, relationships significant at the .05 level of confidence were found, indicating that opinions

TABLE XI

RELATIONSHIPS BETWEEN INSTRUCTOR OPINION AND LENGTH OF TIME EXPOSED TO EXPERIMENTAL GRADING

RELATIONSHIPS TESTED	R*	S/varS**
Length of time vs. safety opinion	.34	1.14
Length of time vs. superiority over ATJ	.66	2.37 (.05)
Length of time vs. value for check flights	.54	2.00 (.05)
Length of time vs. shortening of forms	.43	1.48

* Estimate of Pearson product moment coefficient of correlation from tau value.

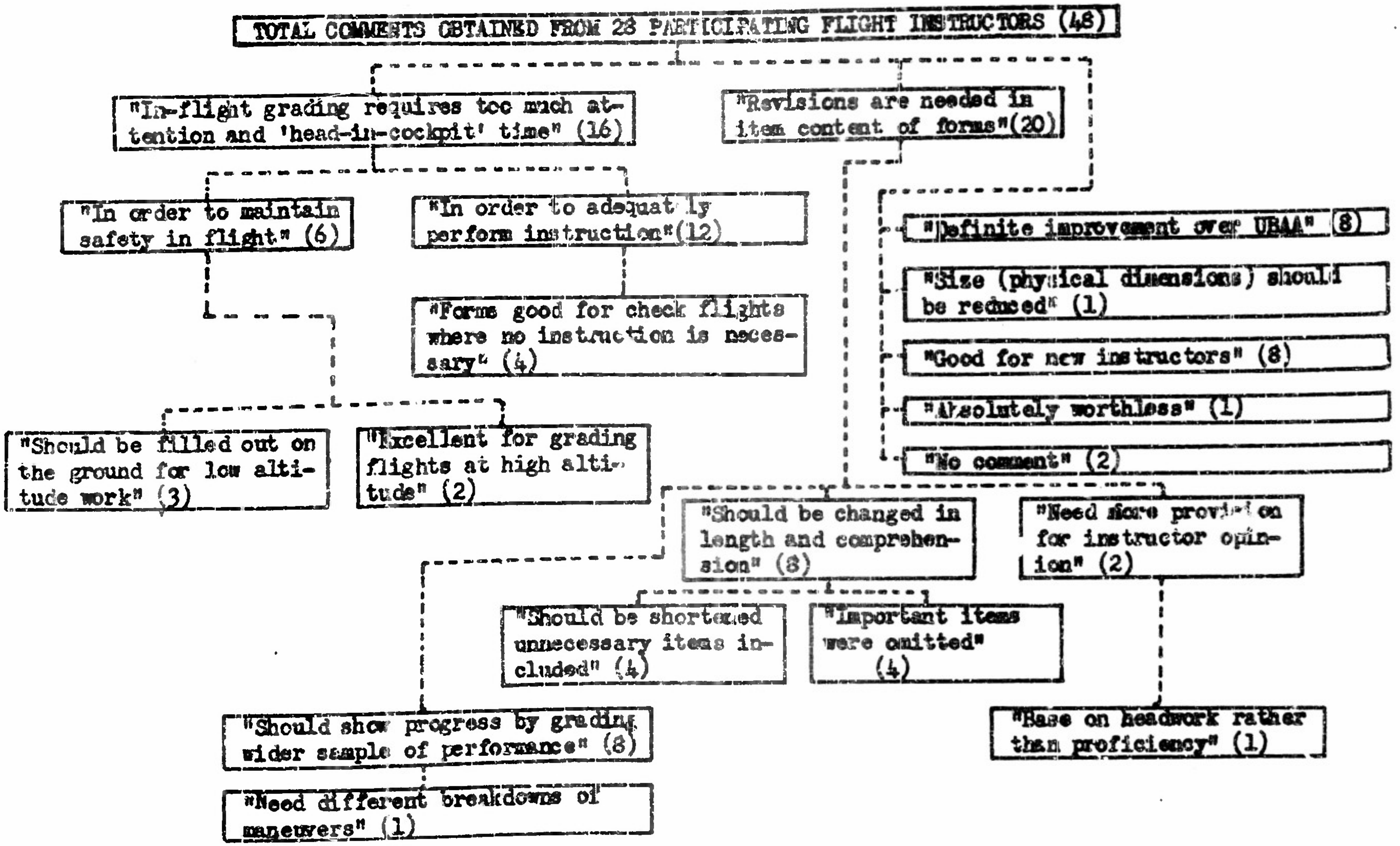
** Positive products minus negative products of a two-fold table divided by the respective variance (6:44).

regarding superiority and value of the experimental grading forms, and regarding the possibility of shortening these forms, is positively related to the number of flights graded. Similar positive relationships are also found between length of time as an experimental instructor and opinions that the forms are superior for specific purposes and that they are of value.

Instructors were given an opportunity to make unstructured comments about the grading forms on these questionnaires. These comments have been summarized in Figure 1 and are of interest at this point. The numbers in parentheses in Figure 1 indicate the number of occurrences of such a statement.

The critical comments were divided generally into two categories; the first of these represents the opinion that in-flight grading requires too much attention, and the second refers to suggested revisions in the content of the form. It is noted in Figure 1 that the idea of in-flight grading requiring 'head-in-cockpit' time was expressed for two reasons differing considerably in nature. One of these considerations was for safety in flight while the other referred to detraction from instruction time. The flow lines in Figure 1 illustrate these, and other opinions, expressed by flight instructors. It must be pointed out to the reader of Figure 1 that the number of occurrences do not total within classifications, since more than one comment may have been made by the same instructor.

FIGURE I



TOTAL COMMENTS OBTAINED FROM 28 PARTICIPATING FLIGHT INSTRUCTORS (48)
"In-flight grading requires too much attention and 'head-in-cockpit' time" (16)
"Revisions are needed in item content of forms"(20)
"In order to maintain safety in flight" (6)
"In order to adequately perform instruction"(12)
"Forms good for check flights where no instruction is necessary" (4)
"Should be filled out on the ground for low altitude work" (3)
"Excellent for grading flights at high altitude" (2)
"Definite improvement over UBAA" (8)
"Size (physical dimensions) should be reduced" (1)
"Good for new instructors" (8)
"Absolutely worthless" (1)
"No comment" (2)
"Should be changed in length and comprehension" (8)
"Need more provision for instructor opinion" (2)
"Should be shortened unnecessary items included" (4)
"Important items were omitted" (4)
"Base on headwork rather than proficiency" (1)
"Should show progress by grading wider sample of performance" (8)
"Need different breakdowns of maneuvers" (1)

CHAPTER VI

DISCUSSION AND CONCLUSIONS

Any conclusions made on the basis of this study must be viewed in light of the limitations existent in the experimental design. Experienced Naval Aviators were used as subjects, and were graded by a select group of instructors, under experimental conditions. Whether, or not, similar results would be obtained in a training unit, composed of cadets and less experienced instructors, can only be suggested.

It appears from the data of this study, however, that flight grading in the Naval Air Basic Training Command could be improved through a basic revision in principle. The general plan of instructor evaluation of a student in terms of quality judgments with respect to other students should be revised in favor of recordings of performance with respect to specific predetermined standards.

It is important to note that the frame of reference used in the Navy ATJ grading form is that of the performance of the "average" student. This consideration is probably the greatest shortcoming of the currently used method. A new instructor actually has no way of knowing the performance of an "average" student. In order to gain clear concept of this hypothetical average against which he is basing his judgments, it is necessary for him to observe a large number of randomly selected performances of each maneuver at different stages of training. Until he has observed such a range of

performance, his concept of the "average" student is necessarily based on the limited sample to which he has been exposed. In many cases, during the first few months, an instructor is, therefore, grading a student against the performance of only the few students he has seen.

There is continuing evidence to support the hypothesis that performance of flight skills varies considerably from flight to flight in such a manner that the usual methods of determining test reliability are not applicable. At the outset of this study, it was felt that the reliability of flight grading could be improved by averaging out these day-to-day variations of student performance in a composite record of several such performances.

Wilcoxon and Johnson (12) found a ride-ride reliability of an objective grading form used on the A19 check flight in the Primary syllabus of this training command in the order of .30. While the split-half reliability of .71 obtained in this study is considerably higher, it still leaves much to be desired in the way of consistency with respect to the usual test standards. Also, this higher reliability can probably be attributed in some measure to the fact that experienced Naval aviators were used in this study as contrasted with Naval air cadets measured by Wilcoxon and Johnson. How much of the reliability increase is due to consistency of day-to-day performance and how much is due to the fact that several performances have been combined is not known. However, an indication of ride-ride reliability is found in the halo analysis where correlation coefficients between grades made on two consecutive flights appeared to

be in the neighborhood of from .35 to .45 for the group under observation. From this it does not appear that Instructors Under Training in the Instructor Basic Training Unit (Flight) are particularly more stable in performance from day to day than are Naval air cadets.

Comparisons of reliabilities between the experimental and the ATJ grading forms revealed that the experimental grades were only slightly more reliable than were ATJ grades. This difference was not significant, but the split-half reliability of .71 for a stage of training probably represents as high a reliability as is possible in a learning situation where day-to-day fluctuations o f performance are known to exist.

Although the grading of attributes appeared to be slightly less reliable than the grading of flight proficiency on the ATJ forms, this factor provided a worthwhile addition to the flight scores obtained on the experimental grading forms. The reliability of experimental grades was equally high with the attribute score added, and was consistently higher in terms of prediction of check flights from daily grades. In view of this, and from instructor opinions, it is felt that the attributes of flight should continue to be judged.

While the ATJ grades were found to be more reliable when used in the experimental group than they were in the control group, this is not the case with prediction of check flights from daily flights. In this measure of validity, it appears that the ATJ grades are slightly superior in the experimental group than they were in the control group. This may be ex-

plained by the fact that the instructors in general became more aware of the critical aspects of flight performance as a result of the tryout of the new grading system and were therefore grading on slightly different criteria than before.

Considerations of safety, especially in an area of air saturation as is found in Pensacola, is of utmost importance in the development of flight grading techniques. While 57 per cent of the instructors felt that these grading forms were a detriment to safety, it was felt by the investigators that the actual safety hazard could be considerably reduced through familiarity with the grading forms through continual use for a longer period of time. Reference is made at this point to the findings of Wilcoxon and Johnson (12) on this question of safety. They found that 65 per cent of the instructors felt that the objective check flight grading form used at Whiting Field was unsafe. The large majority of those instructors, however, felt that the forms would not be a detriment to safety in the air if they were shortened. Considering this point, it is interesting to note that approximately the same percentage of instructors in this study reported herein felt that the experimental grading forms were unsafe, although these forms were only approximately one-fourth as long as the forms used by Wilcoxon and Johnson.

Evidence was found to support the hypothesis that a more favorable attitude toward grading forms of this type are found among instructors who have used the forms for a longer period of time. It is, therefore, important that a thorough period of indoctrination be given instructors in the use of such

forms before basing decisions upon instructor reactions. In the implementation of any new method of grading, especially one which requires more work on the part of the instructor and one which may introduce a need for greater attention to safety considerations, it is reasonable to expect that attitudes would be somewhat unfavorable until such grading became entirely familiar.

The major conclusion resulting from this study centers around the finding that the experimental grading forms were influenced to a distinctly lesser degree by knowledge of past performance. In the halo analysis described in Chapter V above, it was shown that when two consecutive flights were graded by the same instructor, the relationships between these two consecutive flights was significantly higher than when graded by two different instructors, on the Navy ATJ grading forms. In contrast to this, the grades made on two consecutive flights showed little difference attributable to the assignment of instructors when graded by the experimental forms. It is felt that this difference is attributable to the fact that the experimental grading was based on recordings of what the student actually did on the flight.

Such recordings need not be in the form of strictly objective recordings of instrument and dial readings, or of specific recordings of aircraft attitudes, etc. Many of the aspects of a student's performance can be semi-objectively reported in the form of noting whether or not the student performed properly or improperly with respect to specific standards. As noted

in the contents of the grading forms developed in this study, there are many aspects of student flight performance where the student is required to do a specific sub-task. Grading in such cases can be accomplished by merely recording whether or not the task was performed, not how well it was performed.

It is theoretically impossible for grades based on unstandardized "averages" of unknown groups of students to actually reveal the relative quality of the student involved. It is, therefore, felt that in the interest of meaningful evaluation of flight ability, grading procedures be developed on a basis of recordings of exact performance, rather than on subjective judgments. Records based on grades of this type will be more meaningful and will better reflect student quality than those in current use in the Naval Air Basic Training Command.

CHAPTER VII

RECOMMENDATIONS FOR IMPLEMENTATION OF IMPROVED GRADING TECHNIQUES

The two major recommendations which are made on the basis of this research are:

1. That flight grading in the Naval Air Basic Training Command be revised so that it is based upon recordings of actual student performance rather than judgments of quality based on a hypothetical "average".

2. That instructor judgments of attributes such as "mental attitude", "headwork", "reaction toward flight", and "air discipline" be continued as part of any future grading system developed.

In order to carry out a program of revision in grading procedures, several basic steps must be followed. These steps involve the development of in-flight grading forms which differ in content for each flight of each stage within a unit. Such forms, however, would commonly be based on the principle of a detailed report of performance with respect to pre-determined standards. In this way, a new instructor would be able to accurately report the performance of his student with out having to rely on experience for making judgments.

In some cases, it seems feasible that certain portions of a flight grading form might be filled out on the ground, following the flight. Care must be taken in the administration of such a program to see that such reportings are made immediately. One fault of the current method of grading is that several hours, and in some cases several days, elapse before the form is filled out.

Detailed accounts of exactly what the student did will find their greatest use as diagnostic instruments. The weak points of a student are spotlighted, and, in many cases, careful scrutiny of a flight will enable an instructor to discover the underlying causes of a student's weakness.

It is felt that if a program of in-flight grading is established along these lines, following the steps indicated below, a sounder basis may be reached for effective screening and adequate evaluation of training methods.

1. Individual grading forms should be developed to meet the specific needs of each training syllabus.

2. The critical aspects of each maneuver to be graded must be determined by experienced instructors.

3. Standards of performance should be established against which ratings are to be made.

4. These standards must be clearly defined and thoroughly understood by all instructors through a continuing program of indoctrination and standardization.

BIBLIOGRAPHY

1. Bennett, G. K. and others. Flight Jacket Analysis. New York: The Psychological Corporation, Bureau of Medicine and Surgery, Division of Aviation Medicine Report. Contract N 1408-8591 B, May, 1949.

2. Danneskiold, R. D. Evaluation of Student Performance in Operational Flight Trainers. The Psychological Corporation, Special Devices Center Technical Report 999-2-4. Port Washington, L. I., N. Y. 1954

3. Gordon, T. The Development of a Standard Flight-Check for the Airline Transport Rating Based on the Critical Requirements of the Airline Pilots' Job. Civil Aeronautics Administration Division of Research, Report No. 85, Washington, D. C., 1949.

4. Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill Book Company, Inc. 1942.

5. Gulliksen, H. Theory of Mental Tests. New York: John Wiley & Sons, Inc. 1950.

6. Kendall, M. G. Rank Correlation Methods. London: Charles Griffin & Company, Limited. 1948.

7. Lindquist, E. F. Statistical Analysis in Educational Research. Boston: Houghton Mifflin Company. 1940.

8. Mahler, W. R. and Bennett, G. K. Psychological Studies of Advanced Naval Air Training; Analysis of Flight Performance Ratings. The Psychological Corporation, Special Devices Center Technical Report SDC 999-1-2, 1950.

9. Miller, N. E. (Ed.) Psychological Research on Pilot Training. Army Air Forces Aviation Psychology Research Report No. 8, Washington: U. S. Government Printing Office, 1947.

10. Nagay, J. A. The Airline Tryout of the Standard Flight-Check for the Airline Transport Rating. Washington, CAA Division of Research, Report No. 88, December, 1949.

11. National Research Council Committee on Selection and Training of Aircraft Pilots. Standard Check Flight Procedures. Washington, D.C.: CAA Division of Research, Bulletin No. 1, 1942.

12. Wilcoxon, H. C., Johnson, W. and Golan, D. L. The Development and Tryout of Objective Check Flights in Pre-Solo and Basic Instrument Stages of Naval Air Training. The Psychological Corporation and the U. S. Naval School of Aviation Medicine Joint Project Report No. NM 001 058.24.01, 1952.

APPENDIX A

THE NAVY ATJ GRADING FORM

CNATRA Form
ATJ-18-1
PAT

BINDER MARGIN
DO NOT WRITE ABOVE THIS LINE

Navy—PPO CNATRA, Pensacola, Fla.

Only maneuvers which have been introduced prior to or on this flight in accordance with the syllabus shall be graded. Attributes will be graded on every flight. Marks shall be awarded comparatively on the basis of the expected progress toward the established standard.

MANEUVER	Unsat.	Below Average	Average	Above Average
Cockpit Check				
Level Flight				
Turns				
Taxiing				
Take-Off				
Slow Flight				
Transitions				
Landing Pattern				
Stalls				
Spirals				
Landing				
Spin				
Emergencies				
Approaches				
X-Wind Landing Procedure				
Headwork				
Reaction Toward Flt.				
Air Discipline				
Mental Attitude				
Total Marks this hop				
Cumulative Flt. Totals				

COMMENTS

(Check one)
Instructional Flight ☐ Check Flight ☐

(Check one)
Up ☐ Down ☐

Student ______________ Original Class ______________ Flight No. ______________

Date ______________ Training Unit ______________

Instructor's Signature ______________

BASIC PRIMARY—STAGE "A", PRIMARY SOLO

APPENDIX B

THE EXPERIMENTAL GRADING FORMS

TOUCH-AND-GO LANDINGS

(Grade only the first two approaches of type checked below)

1/2 Flap   Full Flap   No Flap

INDICATE THE APPROACHES GRADED ON THIS PAGE:

1st ☐ 2nd ☐ 3rd ☐ 4th ☐ 5th ☐ 6th ☐ 7th ☐ 8th ☐ 9th ☐ 10th ☐

Interval:

Proper [1] [2]   Long [1] [2]   Short [1] [2]

Voice Reports:

	Yes	No	Yes	No
Wheels Down	1	2	1	2
Brakes Pumped	1	2	1	2

Downwind Leg:

Airspeed 80 85 90 95 100   Altitude 400 450 500 550 600

(Indicate first and second landings by a "1" and a "2".)

Approach:

Airspeed at the 90° position 70 75 80 85 90 95

Alignment
- Correct [1] [2]
- Left [1] [2]
- Right [1] [2]

X-Wind Procedure (grade only if these are X-wind landings)
- Proper [1] [2]
- Improper [1] [2]

Straightaway:

Distance
- Proper [1] [2]
- Short [1] [2]
- Long [1] [2]

Transition Altitude
- Proper [1] [2]
- High [1] [2]
- Low [1] [2]

Touchdown:

Contact
- Straight [1] [2]
- Left Skid [1] [2]
- Right Skid [1] [2]

Landing
- 3 Point [1] [2]
- High Stall [1] [2]
- Wheels [1] [2]

Point on Runway
- Within Area [1] [2]
- Outside Area [1] [2]

Stick on Rollout
- Full Back [1] [2]
- Other [1] [2]

Takeoff Nose Attitude:
- Proper [1] [2]
- Low [1] [2]
- High [1] [2]

PsychCorp-SAM 12-2(Rev 3)

GENERAL ASSESSMENT OF ALL TOUCH-AND GO LANDINGS

To be graded on ground immediately after flight

S = Smooth
SE = Slightly Erratic
ONC = Omitted a Necessary Correction

MANNER OF MAKING CORRECTIONS:	S	SE	ONC
M/C Track in Downwind Leg	☐	☐	☐
M/C Altitude and A/S - Downwind Leg	☐	☐	☐
M/C Track in Approach	☐	☐	☐
M/C Altitude and A/S - Approach	☐	☐	☐
M/C Altitude and A/S - Straightaway	☐	☐	☐
M/C Direction during Rollout	☐	☐	☐
M/C Direction during Takeoff	☐	☐	☐
M/C Altitude and A/S during Takeoff and Climbout	☐	☐	☐

	Proper	Slightly Improper
Use of Trim Throughout Hop	☐	☐
Use of Flap Throughout Hop	☐	☐

	No Error	One or More Errors
Use of Prop Pitch	☐	☐

STANDARD FIELD ENTRY

ENTRY INTO TRAFFIC CIRCLE:

Proper ☐
Improper ☐

ALTITUDE IN TRAFFIC CIRCLE:

800 900 1000 1100 1200

AIRSPEED IN TRAFFIC CIRCLE:

110 115 120 125 130

SELECTS BEST TANK:

Yes ☐
No ☐

POWER RETARD, WHEELS DOWN
AND POWER ADDITION:

Proper ☐
Improper ☐

ALTITUDE UNTIL REACHING 95 KTS:

800 900 1000 1100 1200

LOWERS 1/2 FLAP:

Proper ☐
Slightly High ☐
Slightly Low ☐
Below 500' ☐

LANDING XO LIST VOICE REPORT:

Proper ☐
Early or Late ☐
Omits One Item ☐

AIRSPEED IN LETDOWN:

85 90 95 100 105

LETDOWN TRACK:

Proper Distance ☐
Improper Distance ☐

500' TRANSITION:

Smooth ☐
Slightly Rough ☐

TRANSITION ALTITUDE:

400 450 500 550 600

TRANSITION AIRSPEED:

80 85 90 95 100

OVERALL RATING OF ENTRY:

Excellent ☐
Good ☐
Fair ☐
Poor ☐

PsychCorp-SAM 3-2(Rev 2)

STEEP TURNS

LEFT TURN	RIGHT TURN
ENTRY PROCEDURES	
Proper ☐	Proper ☐
Improper ☐	Improper ☐
ANGLE OF BANK (Check both Minimum and Maximum Angle Observed)	
30 35 40 45 50 55 60	30 35 40 45 50 55 60
ALTITUDE CONTROL (Check Maximum Deviation from Base Altitude)	
-200 -100 BA +100 +200	-200 -100 BA +100 +200
USE OF POWER	
Proper ☐	Proper ☐
Improper ☐	Improper ☐
ROLLOUT (Check Deviation from Base Heading)	
-10 -5 BH +5 +10	-10 -5 BH +5 +10
USE OF RUDDERS DURING TURNS	
Excellent ☐	Excellent ☐
Good ☐	Good ☐
Fair ☐	Fair ☐
Poor ☐	Poor ☐

PsychCorp-SAM 4-1(Rev 3)

STALLS

(For each type, grade only the first stall after demonstration)

PROGRESSIVE STALL	STEEP TURN STALL (LEFT)	STEEP TURN STALL (RIGHT)
MADE CLEARING TURN:		
Yes ☐	Yes ☐	Yes ☐
No ☐	No ☐	No ☐
ENTRY AIRSPEED: (Indicate Maximum deviation from proper.)		
−10 −5 0 +5 +10	−10 −5 0 +5 +10	−10 −5 0 +5 +10
NOSE ATTITUDE AT TIME OF STALL:		
Proper ☐	Proper ☐	Proper ☐
Slightly High ☐	Slightly High ☐	Slightly High ☐
Slightly Low ☐	Slightly Low ☐	Slightly Low ☐
WING POSITION AT TIME OF STALL:		
Correct ☐	Correct ☐	Correct ☐
Slightly Off ☐	Slightly Off ☐	Slightly Off ☐
Considerably Off ☐	Considerably Off ☐	Considerably Off ☐
POWER SETTING:		
Proper ☐	Proper ☐	Proper ☐
Slightly Improper ☐	Slightly Improper ☐	Slightly Improper ☐
RECOVERY:		
Excellent ☐	Excellent ☐	Excellent ☐
Good ☐	Good ☐	Good ☐
Fair ☐	Fair ☐	Fair ☐
Poor ☐	Poor ☐	Poor ☐

PsychCorp-SAm 4-2(rev 4)

STALLS

(For each type, grade only the first stall after demonstration)

APPROACH TURN STALL	TRIM TAB STALL	SKIDDED TURN STALL
MADE CLEARING TURN:		
Yes No	Yes No	Yes No
ENTRY AIRSPEED: (Indicate Maximum deviation from proper.)		
-10 -5 0 +5 +10	-10 -5 0 +5 +10	-10 -5 0 +5 +10
NOSE ATTITUDE AT TIME OF STALL:		
Proper Slightly High Slightly Low	Proper Slightly High Slightly Low	Proper Slightly High Slightly Low
WING POSITION AT TIME OF STALL:		
Correct Slightly Off Considerably Off	Correct Slightly Off Considerably Off	Correct Slightly Off Considerably Off
POWER SETTING:		
Proper Slightly Improper	Proper Slightly Improper	Proper Slightly Improper
RECOVERY:		
Excellent Good Fair Poor	Excellent Good Fair Poor	Excellent Good Fair Poor

STALLS

(For each type, grade only the first stall after demonstration)

	NEAR STALL	CLIMBING TURN STALL	POWER-OFF STALL
MADE CLEARING TURN:			
	Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐
ENTRY AIRSPEED: (Indicate Maximum deviation from proper.)			
	-10 -5 0 +5 +10	-10 -5 0 +5 +10	-10 -5 0 +5 +10
NOSE ATTITUDE AT TIME OF STALL:			
	Proper ☐ Slightly High ☐ Slightly Low ☐	Proper ☐ Slightly High ☐ Slightly Low ☐	Proper ☐ Slightly High ☐ Slightly Low ☐
WING POSITION AT TIME OF STALL:			
	Correct ☐ Slightly Off ☐ Considerably Off ☐	Correct ☐ Slightly Off ☐ Considerably Off ☐	Correct ☐ Slightly Off ☐ Considerably Off ☐
POWER SETTING:			
	Proper ☐ Slightly Improper ☐	Proper ☐ Slightly Improper ☐	Proper ☐ Slightly Improper ☐
RECOVERY:			
	Excellent ☐ Good ☐ Fair ☐ Poor ☐	Excellent ☐ Good ☐ Fair ☐ Poor ☐	Excellent ☐ Good ☐ Fair ☐ Poor ☐

PRECISION SPINS

	LEFT SPIN	RIGHT SPIN
ENTRY PROCEDURE:		
Correct	☐	☐
Incorrect	☐	☐
PROPER USE OF CONTROLS DURING ROTATION:		
Yes	☐	☐
No	☐	☐
PROPER USE OF CONTROLS FOR RECOVERY:		
Yes	☐	☐
No	☐	☐
STARTS RECOVERY:		
At 1-1/2 turns	☐	☐
Slightly Early	☐	☐
Slightly Late	☐	☐

AIRSPEED ON RECOVERY:
(Indicate left spin by a "1", and right spin by a "2".)

120 130 140 150 160

SPIRALS

POWER-ON		POWER-OFF	
ENTRY PROCEDURE			
Correct	☐	Correct	☐
Fails to Establish Normal Cruise	☐	Fails to Establish Normal Cruise	☐
Fails to Establish Proper Glide	☐	Fails to Establish Proper Glide	☐

ANGLE OF BANK
(Indicate maximum deviation throughout)

Power-on: 30 — 45 — 60

Power-off: 30 — 45 — 60

AIRSPEED
(Indicate maximum deviation throughout)

Power-on: 110 — 115 — 120 — 125 — 130

Power-off: 85 — 90 — 95 — 100 — 105

RECOVERY PROCEDURES

POWER-ON		POWER-OFF	
Correct	☐	Correct	☐
Incorrect	☐	Incorrect	☐

SLOW FLIGHT

ENTRY:	Proper ☐	One or More Errors ☐		
THROTTLE:	Excellent ☐	Good ☐	Fair ☐	Poor ☐
NOSE ATTITUDE:	Excellent ☐	Good ☐	Fair ☐	Poor ☐
RECOVERY:	Proper ☐	One or More Errors ☐		

MAXIMUM HEADING DEVIATION: −30 — −20 — −10 — BH — +10 — +20 — +30

MAXIMUM ALTITUDE DEVIATION: −150 — −100 — −50 — BA — +50 — +100 — +150

HIGH ALTITUDE EMERGENCY

INITIAL PROCEDURES: Correct ☐
One or More Errors ☐

VOICE REPORT: Correct ☐
One or More Errors ☐

PATTERN (to wave-off position):
Excellent ☐ Good ☐ Fair ☐ Poor ☐

AT 1500' ALTITUDE:

Correct Distance ☐ On #1 Position ☐
Slightly Close ☐ Slightly off #1 Position ☐
Slightly Wide ☐

AT 500' ALTITUDE

Voice Report Correct Distance ☐
Yes ☐ Would Overshoot ☐
No ☐ Would Undershoot ☐

WAVE-OFF PROCEDURE:

Correct ☐
Incorrect ☐

PsychCorp-SAM 5-4(Rev 3)

SMALL FIELD PROCEDURES AND LOW ALTITUDE EMERGENCY

	1st SHOT	2nd SHOT
POSITION & DISTANCE AT CUT:		
Correctly on	☐	☐
Ahead or Behind	☐	☐
Close or Wide	☐	☐

ALTITUDE AT CUT:

900 — 950 — 1000 — 1050 — 1100

AIRSPEED AT CUT:

110 — 115 — 120 — 125 — 130

ALTITUDE AT 95 KNOTS:

900 — 950 — 1000 — 1050 — 1100

	1st SHOT	2nd SHOT
PROP PITCH:		
Yes	☐	☐
No	☐	☐
WHEELS CHECK:		
Yes	☐	☐
No	☐	☐
GLIDE PATTERN:		
Excellent	☐	☐
Good	☐	☐
Fair	☐	☐
Poor	☐	☐

AIRSPEED UNTIL FLAPS DOWN:

85 — 90 — 95 — 100 — 105

	1st SHOT	2nd SHOT
LANDING TRANSITION:		
Excellent	☐	☐
Good	☐	☐
Fair	☐	☐
Poor	☐	☐

	1st SHOT	2nd SHOT
POINT OF TOUCHDOWN:		
First 1/3	☐	☐
Beyond	☐	☐
Wave-Off: (Student error)	☐	☐
CLIMB OUT PROCEDURES:		
Correct	☐	☐
Missed one or more items	☐	☐
CLIMB OUT PATTERN:		
Excellent	☐	☐
Good	☐	☐
Fair	☐	☐
Poor	☐	☐
TRANSITION TO NORMAL CRUISE:		
Proper	☐	☐
Improper	☐	☐

LOW ALTITUDE EMERGENCY

GEAR UP CHECK

Yes ☐ No ☐

PROP PITCH LOW:

Yes ☐ No ☐

SELECTION OF LANDING SITE:

Excellent ☐ Fair ☐
Good ☐ Poor ☐

USE OF FLAP:

Proper ☐ Improper ☐

WAVE OFF PATTERN:

Excellent ☐ Fair ☐
Good ☐ Poor ☐

PsychCorp-SAM 8-2(Rev.4)